Historiallinen korpuslingvistiikka

Vartiainen, Turo Anssi Kalevi

Suomalaisen Kirjallisuuden Seura
2020

Vartiainen , T A K & Säily , T 2020 , Historiallinen korpuslingvistiikka . julkaisussa M Luodonpää-Manni , M Hamunen , R Konstenius , M Miestamo , U Nikanne & K Sinnemäki (toim) , Kielentutkimuksen menetelmiä I-IV . Vuosikerta. II , Suomalaisen Kirjallisuuden Seuran Toimituksia , Nro 1457 , Suomalaisen Kirjallisuuden Seura , Helsinki , Sivut 525-556 .

http://hdl.handle.net/10138/323311

Luku 11 Historiallinen korpuslingvistiikka

Turo Vartiainen
https://orcid.org/0000-0002-4760-750X

Tanja Säily
https://orcid.org/0000-0003-4407-8929

Mikä?

Historiallinen korpuslingvistiikka tarjoaa menetelmän eri kielten historian ja muutoksen tutkimiseen laajojen, systemaattisesti koottujen elektronisten tekstikokoelmien avulla. Menetelmässä yhdistyvät laadullinen ja määrällinen näkökulma, ja siinä korostuu tutkittavan aineiston tarkka tuntemus. Historiallisen korpuslingvistiikan menetelmin voidaan saada uutta tietoa paitsi kielestä järjestelmänä myös ihmisyhteisöjen kielikäytäntöjen vaihtelusta ja muutoksesta.

Katso myös:

Luku 7 Laadullinen aineistopohjainen kielentutkimus

Luku 9 Korpusaineistot

Luku 10 Määrällinen korpuslingvistiikka

Luku 15 Historiallinen ja vertaileva kielentutkimus

1. Johdanto

Kielen varhaisvaiheiden tutkimus on lähtökohtaisesti aina aineistopohjaista. Oli kyseessä sitten yksittäisen kielen muutos ajassa tai kokonaisen kielikunnan varhaisimman oletetun vaiheen, protokielen, rekonstruktio, tutkimus pohjautuu aina olemassa oleviin, usein sattumalta nykyaikaan saakka säilyneisiin teksteihin. Historiallinen korpuslingvistiikka on menetelmä, jonka avulla voidaan systemaattisesti tutkia eri kielten varhaisempia vaiheita monesta eri näkökulmasta. Kyseessä on empiirinen metodologia, jossa tutkittavan kieliaineiston niukkuus voi toisinaan asettaa tutkimukselle tiettyjä rajoituksia ja haasteita mutta tarjoaa samalla mahdollisuuden tarkastella paitsi hyvin yksityiskohtaisia kielioppiin, kielenmuutokseen ja kielen käyttöön liittyviä kysymyksiä myös abstraktimpia, perustavanlaatuisia kysymyksiä kielestä dynaamisena, ajassa muuttuvana ilmiönä. Yksityiskohtaisemmat, mikrotason tutkimuskysymykset, jotka kiinnostavat historiallisia korpuslingvistejä, voivat koskea esimerkiksi kielen taivutusjärjestelmän, apuverbirakenteiden tai sanajärjestyksen vaihtelua ja muutosta. Laajempia, makrotason teemoja, joihin on viime aikoina alettu kiinnittää yhä tarkempaa huomiota, ovat esimerkiksi seuraavat kysymykset:

1) Mitkä kielensisäiset tekijät edistävät kielenmuutosta ja millä tavoin?
2) Mitkä kielenulkoiset tekijät, kuten eri kieltä puhuvien ihmisten väliset kontaktit ja kieliyhteisöjen rakenne, vaikuttavat kielenmuutokseen?
3) Millaiset ihmiset tai ihmisryhmät omaksuvat erityisen nopeasti uusia tapoja käyttää kieltä ja tällä tavoin edesauttavat kielenmuutoksen leviämistä? Ketkä ovat konservatiivisia kielenkäyttäjiä?

Kuten näistä teemoista käy ilmi, suuri osa historiallisista korpuslingvisteistä on kiinnostunut nimenomaan siitä, millä tavoin ja missä olosuhteissa kieli muuttuu ajan saatossa. Tutkimuksen näkökulma on siis usein **diakroninen**, mutta kielten varhaisempia vaiheita on mahdollista

tarkastella myös **synkronisesta** näkökulmasta, ikään kuin läpileikkauksena tiettynä aikakautena puhutusta kielestä. Oli näkökulma kumpi hyvänsä, tutkija joutuu työssään ottamaan huomioon monenlaisia niin aineiston edustavuuteen kuin korpuslingvistiseen menetelmäänkin liittyviä asioita, jotka vaikuttavat siihen, millaisia kysymyksiä korpusaineiston avulla on ensinnäkin mahdollista tutkia, sekä siihen, millaisia johtopäätöksiä tulosten pohjalta voidaan esittää. Tutkijan vastuu on erityisen suuri silloin, kun käytössä on vain niukka ja tyylikirjoltaan suppea tai epäyhtenäinen aineisto, jonka perusteella on hankala tarkastella kielenmuutokseen liittyviä yleisluontoisia hypoteeseja (ks. luku 2).

Yksi historiallisen korpuslingvistiikan suurimmista eduista on sen tarjoama mahdollisuus kielen ilmiöiden yleisyyden eli frekvenssin tarkasteluun. Frekvenssitietojen avulla tutkija voi tehdä tarkkoja havaintoja ajassa vähitellen tapahtuvista muutoksista sekä tarkastella sellaisia kielen ilmiöitä, jotka eivät ilmene kategorisina joko–tai-jakoina vaan pikemminkin kielenkäytössä esiintyvinä tendensseinä, joiden ajallinen muutos voi lopulta johtaa koko kielioppijärjestelmän muuttumiseen. Koska korpusaineistossa on aina myös sattumanvaraista vaihtelua, tutkijat hyödyntävät tällaisten kielenmuutosten tutkimuksessa usein tilastollisia menetelmiä, joiden avulla voidaan esittää todennäköisyys sille, onko esimerkiksi kahtena eri aikakautena havaittavassa kielenmuutoksessa kyse todellisesta kielioppijärjestelmän muutoksesta vai onko kyse vain kielenkäyttöön luonnollisesti kuuluvasta vaihtelusta. Esittelemme joitain historiallisessa korpustutkimuksessa yleisesti käytettyjä tilastollisia menetelmiä myöhemmin tässä artikkelissa. Tässä yhteydessä on syytä todeta, että tilastollisten menetelmien käyttö ei historiallisessa korpuslingvistiikassa ole pakollinen osa tutkimusta. On mahdollista, että kielenmuutos etenee ajassa niin selkeästi, että tulokset ikään kuin ”puhuvat puolestaan”. Tilastollisia menetelmiä ei kuitenkaan ole syytä karttaa tai pelätä: tutkijoiden käytössä on nykyään monia helppokäyttöisiä tilastollisia testejä, jotka voivat tarjota aineistosta arvokasta lisätietoa.

Monet historiallista korpuslingvistiikkaa koskevista eduista, haasteista ja rajoitteista koskevat korpuslingvistiikkaa yleisemminkin (ks. Korpusaineistot tässä kirjassa). Nykykielen tutkimuksesta poiketen kielen varhaisempien vaiheiden tutkijat eivät kuitenkaan voi täydentää

korpusaineiston kautta saamaansa käsitystä kielen rakenteesta ja sen muutoksesta konsultoimalla kielen puhujia. Tätä voidaan pitää sekä etuna että haittana: yhtäältä on selvää, että kielenpuhujien intuitiot ja näkemykset omasta kielestään voivat tarjota arvokasta tietoa. Toisaalta on myös selvää, että puhujien intuitiot eivät aina pidä paikkaansa ja että heidän käsityksensä niin omasta kuin muidenkin ihmisten kielenkäytöstä voivat olla virheellisiä. Tästä näkökulmasta historiallinen korpuslingvistiikka näyttäytyy erityisen vahvasti aineistopohjaisena menetelmänä. Menetelmän aineistolähtöisyys ei kuitenkaan tarkoita sitä, etteivätkö historialliset korpuslingvistit voisi selittää tuloksiaan esimerkiksi kognitiivisesta tai sosiolingvistisestä näkökulmasta: vaikka tutkijoilla ei ole pääsyä menneinä aikoina eläneiden ihmisten ajatuksiin, psykologisiin prosesseihin ja käyttäytymismalleihin (esimerkiksi psykologisten testien, kielitajun, kyselytutkimusten tai havainnoinnin avulla), on oletettavaa, että esimerkiksi ihmisen aivotoiminta on biologisesta näkökulmasta ollut samanlaista koko kirjoitetun, n. 5 000 vuoden päähän yltävän historian ajan. Myös ihmisyhteisöjen ja ihmisten sosiaalisten roolien tarkka tutkimus ja ymmärtäminen tarjoavat mahdollisuuden selittää varhaisempien kieliyhteisöjen kielenkäyttöä ja sitä kautta kielenmuutoksen ulkoisia tekijöitä.

Historiallisen korpusaineiston avulla voidaan myös tarkastella ihmisen kognitioon ja kielenmuutokseen liittyviä yleisiä prosesseja. Yksi varsin paljon tutkittu esimerkki tällaisesta prosessista on merkitysten ns. **subjektifikaatio**, mikä tarkoittaa sitä, että ihmisillä on tapana alkaa käyttää sanoja ja rakenteita yhä enemmän omien mielipiteidensä, asenteidensa ja olettamustensa ilmaisuun – merkitykset ikään kuin hivuttautuvat lähemmäksi kielenpuhujaa, kielenkäytön subjektia (Benveniste 1971 [1958]: *le sujet de l'énonciation*). Esimerkiksi englannin kielessä yleisesti käytettyä *be going to* -rakennetta käytettiin alun perin ilmaisemaan vain fyysistä liikettä (esim. *She's going to the market*), kun taas nykyenglannissa rakennetta käytetään yleisemmin ilmaisemaan asioita, joita kielenpuhuja joko olettaa tapahtuvan tulevaisuudessa tai joiden tapahtumisesta hän on epävarma (esim. *It's going to rain*; *Are you going to come to the party?*). Vaikka tällaisia havaintoja kielenmuutoksesta ja sen luonteesta on mahdollista tehdä myös ilman korpusaineistoja, korpuslingvistiikka

tarjoaa monipuolisen menetelmän, jonka avulla voidaan hyvinkin tarkasti ajoittaa kielessä tapahtuvia muutoksia sekä etsiä niitä kielenkäytön kontekseja, jotka ovat olleet erityisen suotuisia muutoksen etenemiselle.

On syytä painottaa, että historiallinen korpuslingvistiikka on menetelmänä täysin riippumaton tutkittavasta kielestä – mitä tahansa kieltä, josta on koottu historiallinen korpus, on mahdollista tutkia historiallisen korpuslingvistiikan menetelmin, eikä ole lainkaan epätavallista, että tutkija itse koostaa korpuksen tutkimuksen kohteena olevasta kielestä. Koska tämän artikkelin kirjoittajien tausta on kuitenkin englannin historiallisessa kielitieteessä ja koska menetelmä on hyvin yleinen englannin kielen tutkimuksessa, jossa sillä on myös erityisen pitkä historia, artikkelissa käytetyt esimerkit ovat pääasiassa englannin kielestä.

2. Metodin historiaa ja taustaa

Historiallisen korpuslingvistiikan voidaan yleisesti katsoa pohjautuvan aineistoperäiseen kielentutkimukseen, jota on tehty vuosisatojen ajan. Uuden ajan Länsi-Euroopassa tutkijoiden mielenkiinnon kohteina olivat etenkin klassiset kielet, latina ja klassinen kreikka, joiden pohjalta alettiin jo 1800-luvulla koostaa kattavia tekstikokoelmia. Näistä kokoelmista on niiden ensi vaiheista lähtien käytetty nimitystä **korpus**, mutta on syytä painottaa, että *korpus*-termiä ei tuolloin käytetty sanan nykyisessä, teknisessä merkityksessä. Esimerkiksi Corpus Inscriptionum Latinarum -projektin pyrkimyksenä on ollut koota kaikki Rooman valtakunnan alueella laaditut latinankieliset piirtokirjoitukset yhteen kriittiseen editioon riippumatta niiden kirjoitusajasta, tarkoituksesta tai kirjoituksen alkuperästä. Nykymerkityksessään *korpuksella* tarkoitetaan kuitenkin tyypillisesti tarkoin harkittujen periaatteiden mukaisesti koottua digitaalista kieliaineistoa: tietokoneluettavaa tekstikokoelmaa, joka sisältää ennalta määritellyssä ja tarkoin harkitussa suhteessa erityyppisiä tekstejä, joiden voidaan katsoa riittävässä laajuudessa edustavan kieltä sen eri käyttöfunktioissa ja rekistereissä (McEnery & Hardie 2012). Ajatuksena on

se, että ihmiset käyttävät kieltä hyvin eri tavoin erilaisissa konteksteissa, ja jos kielitieteellinen korpus sisältää liian suppean valikoiman tekstejä – vaikkapa ainoastaan sanomalehtiartikkeleita –, jää tutkimuksen tavoittamattomiin kaikki se kielen käyttöön liittyvä vaihtelu, joka nousee esiin esimerkiksi epämuodollisissa keskusteluissa, sähköposteissa, blogikirjoituksissa tai kaunokirjallisuudessa. On toki mahdollista, että korpus sisältää tekstejä vain yhdestä genrestä (ks. alla). Tällöin tutkijan on kuitenkin varottava esittämästä liian yleisiä väitteitä kielestä aineistonsa perusteella: kielenmuutokset voivat näkyä eri genreissä eri aikoina, ja jotkin kielen rakenteet voivat olla hyvin yleisiä yhdessä genressä ja äärimmäisen harvinaisia toisessa. Esimerkiksi yksikön ensimmäisen ja toisen persoonan pronominit ovat erittäin yleisiä epämuodollisissa keskusteluissa, mutta niitä ei käytetä juuri lainkaan akateemisissa artikkeleissa.

Oli kyseessä sitten useammasta genrestä koostuva **yleiskorpus** tai vain yhden genren käsittävä korpus, korpuslingvistien tavoitteena on aina tehdä korpuksista mahdollisimman **edustavia**, niin että niiden voidaan ajatella riittävässä määrin kuvaavan kieltä sellaisena kuin se tietyn kieliyhteisön moninaisissa kielenkäyttötilanteissa esiintyy (ks. Korpusaineistot tk.). Toinen, korpuksen edustavuuteen liittyvä ja aivan yhtä tärkeä pyrkimys on saavuttaa **tasapaino** korpuksessa esiintyvien tekstien välillä. Yleiskorpuksessa voidaan esimerkiksi pyrkiä siihen, että eri tekstilajeja, kuten akateemisia artikkeleita, aikakauslehtiä, blogitekstejä ja epämuodollista keskustelua, on korpuksessa suhteellisesti yhtä paljon. Korpuksen tasapainoa voidaan myös parantaa ottamalla huomioon vaikkapa tekstien kirjoittajien sukupuoli ja ikä, jolloin korpusaineiston avulla voidaan tutkia myös sosiolingvistisiä kysymyksiä, kuten miesten ja naisten kielenkäyttöön liittyviä eroja. Korpuksen edustavuuteen ja tasapainoon liittyvät kysymykset koskevat niin synkronisia kuin diakronisia korpuksia, mutta historialliseen korpuslingvistiikkaan ja varsinkin diakroniseen tutkimukseen olennaisesti kuuluva aineistojen välinen vertailu korostaa entisestään näiden metodologisten perusperiaatteiden tärkeyttä.

Vaatimukset korpuksen edustavuudesta ja tasapainosta otettiin huomioon jo 1960-luvun alussa, kun ensimmäistä nykyaikaista

amerikanenglannin korpusta alettiin koota Brownin yliopistossa Yhdysvalloissa. Vuonna 1964 julkaistu ja yhä tutkijoiden käyttämä Brown Corpus[1] koostuu sanomalehtiteksteistä sekä kauno- ja tietokirjallisuudesta, ja sen sisältämät tekstit on luokiteltu yksityiskohtaisesti alakategorioihin: korpus sisältää n. 2 000 sanan otteita esimerkiksi scifi- ja seikkailukirjoista, elämäkerroista ja sanomalehtien pääkirjoituksista (ks. Francis ja Kučera 1979; ks. myös Korpusaineistot tk.). Kiinnostus kielen historiaan ja sen muutokseen johti 1990-luvun alussa uuden, samoilla periaatteilla toteutetun korpuksen kokoamiseen. Freiburg-Brown- eli Frown-korpusta[2] voidaan käyttää kolmekymmentä vuotta vanhemman Brown-korpuksen vertailukorpuksena. Sittemmin Brown-korpusperhettä on täydennetty 1960- ja 1990-luvun brittienglannin korpuksilla (Lancaster-Oslo-Bergen Corpus eli LOB[3]; Freiburg-LOB Corpus eli FLOB[4]) sekä kahdella 1930-luvun britti- ja amerikanenglannin korpuksella (BLOB-1931[5], The 1930s Brown Corpus eli B-Brown[6]). Koska Brown-perheen korpukset vastaavat niin hyvin toisiaan rakenteellisesti, ne tarjoavat erinomaisen mahdollisuuden tarkastella kielenmuutosta noin 60 vuoden aikavälillä sekä britti- että amerikanenglannissa.

Brown- ja Frown-korpukset noudattavat vaatimuksia korpuksen tasapainosta ja edustavuudesta siinä määrin kuin huoliteltu yleiskieli voi ylipäätään antaa tarkan kuvan kielestä, sen vaihtelusta ja muutoksesta. Näitä korpuksen kokoamista koskevia yleisiä periaatteita on kuitenkin sitä hankalampi noudattaa, mitä kauemmas taaksepäin kielen historiassa menemme: vanhoja tekstejä ei yksinkertaisesti ole aina säilynyt riittävästi, jotta valinnanvaraa olisi. Varhaisten kielivaiheiden tutkijat joutuvatkin usein työskentelemään sellaisen aineiston parissa, joka on säilynyt osin sattumalta ja osin taas siksi, että teksti on syystä tai toisesta joko koettu yhteiskunnallisesti tärkeäksi ja säilyttämisen arvoiseksi tai tuotettu olosuhteissa, jotka ovat jostain syystä edesauttaneet sen säilymistä. Valtaapitävien tekstit ovat kautta historian säilyneet huomattavasti todennäköisemmin kuin tavallisen kansan kirjoitukset, ja tämä näkyy myös monen historiallisen korpuksen tekstivalikoimassa: lakitekstit, luostarisäännöt ja raamatunkäännökset ovat yleensä varsin hyvin edustettuina, kun taas yksityishenkilöiden kirjoittamat kirjeet, päiväkirjat tai muut arkielämää koskevat tekstit ovat hyvin harvinaisia.

Myös kielen murrevaihtelu voi asettaa tutkimukselle omat haasteensa. Esimerkiksi valtaosa muinaisenglannin ajalta (n. 700–1150 jaa.) säilyneistä teksteistä on kirjoitettu Etelä-Englannissa sijainneen Wessexin kuningaskunnan murteella. Koska muinaisenglanniksi kirjoitettuja tekstejä on säilynyt nykypäivään saakka vain noin kolme miljoonaa sanaa, on tämä materiaali otettu kokonaisuudessaan osaksi eräitä englannin kielen historiallisia korpuksia. Tällöin ongelmaksi muodostuu se, että valtaosa englannin kielen seuraavan vaiheen, keskienglannin (n. 1150–1500 jaa.), ajalta säilyneistä teksteistä on peräisin Wessexiä pohjoisemmilta Itä- ja Länsi-Midlandsin alueilta. Tällaisia korpuksia työssään käyttävä kielenmuutoksen tutkija joutuukin pitämään mielessään, että aineistossa havaittava muutos voi selittyä sillä, että eri murteiden suhteellinen osuus ei ole kaikkina korpuksen edustamina aikakausina tasapainossa. Tilanne voi pahimmillaan vastata sitä, että suomen kielen lähihistoriaa tutkittaisiin vertailemalla 1800-luvun Kuopion murretta 2000-luvun Turun murteeseen ja tämän perusteella tehtäisiin yleisluontoisia, koko suomen kieltä koskevia johtopäätöksiä.

Yksi varhaisimmista, ja samalla kunnianhimoisimmista, historiallisista korpuksista on Helsinki Corpus of English Texts (HC)[7], jonka kokoaminen aloitettiin 1980-luvulla Helsingin yliopistossa professori Matti Rissasen johdolla. Helsinki Corpus julkaistiin vuonna 1991, ja sen pohjalta on sittemmin kirjoitettu satoja tieteellisiä artikkeleita, jotka valaisevat englannin kielessä tapahtunutta muutosta muinaisenglannista 1700-luvulle. Korpuksen vahvuutena on sen kokoamistyössä noudatettu huolellisuus: eri aikakausien tekstit on luokiteltu yksityiskohtaisiin kategorioihin, ja pyrkimyksenä on ollut esittää englannin kielen yli tuhatvuotinen jatkumo niin, ettei kokonaiskuva liiaksi hämärtyisi murrevaihtelun tai tekstilajeissa tapahtuvien muutosten vuoksi. Helsinki Corpus koottiin aikana, jolloin yhteiskunnan tietokoneistuminen oli vasta alkanut. Tämä näkyy omalta osaltaan myös korpuksen koossa: korpus käsittää kaikkinensa n. 1,5 miljoonaa sanaa, mikä riittää vain suhteellisen yleisten kielioppi-ilmiöiden ja sanastonmuutoksen tutkimiseen. Toisaalta se, mikä koossa hävitään, voitetaan tarkkuudessa. Viime vuosina julkaistut ns. ”megakorpukset”, kuten yli 400 miljoonaa sanaa vuosilta 1810–2009 käsittävä Corpus of Historical American English

(COHA)[8], ovat kokoamisperiaatteiltaan melko suurpiirteisiä: tekstien kategorisointi on toteutettu karkeammin, ja koska korpuksen suuren koon vuoksi ei tutkijoilla ole ollut mahdollisuutta tarkastaa tekstejä manuaalisesti, on korpuksessa jonkin verran niin tekstien ajoitusta, kirjoitusmuotoja kuin kielen varianttiakin koskevia virheitä.

Riippumatta käytetystä korpuksesta kielentutkijan vastuulla on aina tuntea aineistonsa mahdollisimman tarkasti, niin että aineiston epätasapaino tai eri tekstilajien konventioissa tapahtuvat muutokset eivät johda vääriin tulkintoihin. Esimerkiksi kaikki historialliset tekstilajit, kuten pyhimyselämäkerrat tai erilaiset pamfletit, eivät sellaisinaan ole säilyneet nykypäivään, ja tämä on hyvä pitää mielessä korpusaineistoa tutkittaessa (ks. myös Heikkinen, Voutilainen, Lauerma, Tiililä & Lounela 2012). Toisaalta voi syntyä uusia genrejä: esimerkiksi sanomalehdet kehittyivät pääasiassa 1700-luvulta lähtien, ja internetin yleistymisen myötä on viime vuosikymmeninä kehittynyt monia uusia tekstilajeja. Myös tekstilajeja koskevat käytänteet muuttuvat usein, ja tällä voi olla vaikutusta korpusaineistosta saatuihin tietoihin: esimerkiksi yksityiskirjeiden muodollisuus ja siihen liittyvät kielelliset piirteet sekä tieteellisen tekstin syntaktinen kompleksisuus ovat muuttuneet ajan saatossa. Kuitenkaan nämä esimerkiksi pronominien ja passiivirakenteiden frekvenssissä näkyvät muutokset eivät ole todiste kielenmuutoksesta vaan selittyvät yksittäisten tekstilajien esitystavoilla ja niihin liittyvillä muutoksilla.

Historiallisen korpuslingvistiikan aineistojen kehityksessä on nähtävissä useita eri tendenssejä. Toisaalta on saatavilla yhä suurempia korpuksia ja tietokantoja, kuten Early English Books Online (EEBO) ja Kansalliskirjaston lehtikokoelma, jotka sisältävät massiivisen määrän painettuja tekstejä ja joiden avulla voidaan tutkia entistä harvinaisempia muutoksia.[9] Toisaalta myös pienempiä korpuksia kootaan edelleen, mutta niiden annotointi on aiempaa kehittyneempää: puhutaan ns. ”rikkaasta datasta”, josta voi löytyä genre- ja sanaluokka-annotaation lisäksi vaikkapa kirjoittajien taustatietoja, tarkkoja kuvauksia tekstin asettelusta ja käytetyistä käsialoista tai kirjasintyypeistä sekä valokuvia alkuperäisteksteistä. Siinä missä varhaiset miljoonan sanan korpukset pyrkivät edustamaan kieltä yleensä ja sisälsivät useita genrejä, nykyisin pienet korpukset keskittyvät usein vain yhteen genreen, kuten kirjeisiin (esim.

Corpora of Early English Correspondence, CEEC)[10] tai jopa yhden kirjoittajan teksteihin. Tällainen korpus voidaan koota myös osana opinnäytetyötä, jossa voidaan sitten käyttää korpusta tutkimukseen (esim. Marttila 2014). Sekä isoa että rikasta dataa käytetään yhä enemmän digitaalisissa ihmistieteissä yleensä, ja samoja aineistoja voivat hyödyntää niin historioitsija, lingvisti kuin kirjallisuudentutkijakin.

3. Kuinka historiallista korpuslingvistiikkaa harjoitetaan

3.1. Tutkimuksen kulku

3.1.1. AINEISTON HANKKIMINEN JA KÄSITTELY

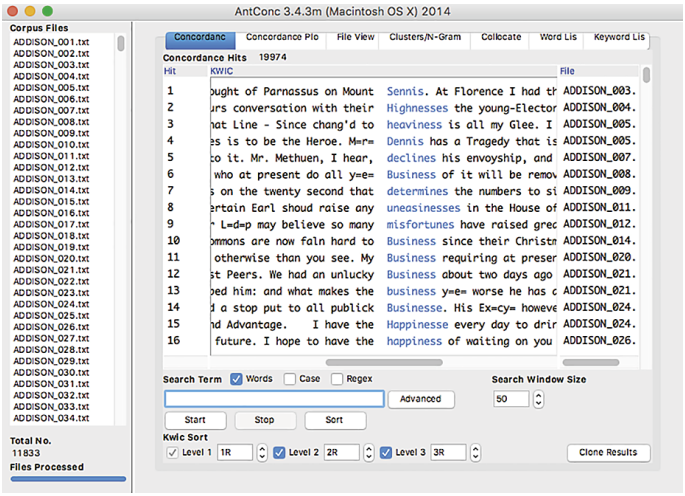
Historiallisessa korpuslingvistiikassa tutkimus aloitetaan tyypillisesti tekemällä aineistosta hakuja, joilla tavoitetaan se kielenaines, jonka muutosta halutaan tutkia. Tarvitaan siis sopiva korpus sekä jokin hakutyökalu. Historialliset korpukset voidaan lajitella kahteen luokkaan: Alun perin diakronisiksi kootuista korpuksista (esim. Helsinki Corpus ja Vanhan kirjasuomen korpus)[11] löytyy tekstejä enemmän tai vähemmän tasaisesti koko niiden kattamalta aikaperiodilta. Toisaalta on myös synkronisia korpuksia, jotka edustavat vain yhtä ajanhetkeä mutta joita on myöhemmin laajennettu kokoamalla vastaava korpus eri ajanhetkeltä. Jälkimmäistä tyyppiä edustaa aiemmin mainittu Brown-korpusperhe (ks. Korpusaineistot tk.).

Korpuksia on saatavilla monista eri lähteistä. Esimerkiksi Kielipankkiin[12] on koottu laaja kokoelma erikielisiä korpusaineistoja. Kielipankki on osa FIN-CLARIN-konsortiota, joka puolestaan kuuluu eurooppalaiseen CLARIN-organisaatioon, jonka tarkoituksena on kehittää ja tarjota kieliaineistoja sekä näiden aineistojen käsittelyyn sopivia työkaluja. Korpuksia voi etsiä myös metadatapalveluista. Historialliselle korpuslingvistiikalle ei vielä ole omaa palvelua, mutta yleisimpiin palveluihin lukeutuvat muiden muassa META-SHARE, CLARINin Language Resource

Inventory, Linguistic Data Consortiumin luettelo sekä englannin kielen osalta Corpus Resource Database (CoRD).[13] Korpukset eivät useinkaan ole täysin vapaasti saatavilla mm. tekijänoikeusongelmien takia, ja osa niistä on maksullisia. Yliopistojen kielten laitoksilla on tyypillisesti lisenssejä yleisimpien korpusaineistojen käyttöön, joten ennen lisenssimaksun maksamista kannattaa tarkistaa Kielipankin lisäksi oman yliopiston aineistokokoelmat.

Korpushakutyökalu tulee usein korpuksen mukana: Esimerkiksi COHA-korpusta voi käyttää sille räätälöidyllä web-käyttöliittymällä. Korpuksen kokoaja, Mark Davies, on laatinut muitakin historiallisia korpuksia, jotka ovat kaikki käytettävissä vastaavan nettikäyttöliittymän avulla (esim. espanjan- ja portugalinkieliset Corpus del Español ja Corpus do Português).[14] Suomen web-pohjaisista palveluista mainittakoon Kielipankin Korp[15] (ks. myös Korpusaineistot tk.). Korp-palveluun on sovitettu useita koti- ja ulkomaisia korpuksia, joista osa on diakronisia ja osa synkronisia. Kansainvälisesti yleinen web-pohjainen käyttöliittymä on CQPweb; esimerkiksi Lancasterin yliopiston CQPweb-palvelimelle on asennettu laaja kirjo etenkin englanninkielisiä korpuksia.[16]

Joitakin korpuksia voi ladata myös omalle koneelle ja käyttää erillisen konkordanssiohjelman avulla. Tällaisia ohjelmia ovat esimerkiksi maksullinen WordSmith Tools ja ilmainen AntConc (kuva 1), joiden käyttöön on internetissä hyvät ohjeet.[17] Konkordanssiohjelman perustoimintoihin kuuluu haku, jonka tulokset näytetään yleensä kuvan 1 tapaisena konkordanssilistauksena. Kukin hakutulos näytetään omalla rivillään, jonka keskellä on haettu sana, fraasi tai sanan osa, ja ympärillä konteksti, jossa hakutulos esiintyy. Ohjelmointitaitoiset voivat tehdä korpushakuja myös ohjelmointikielellä, vaikkapa R-ympäristössä.[18] Tämän lähestymistavan etuja on se, että tutkija ei ole sidottu valmiissa korpusohjelmissa oleviin toimintoihin, sekä se, että tutkimuksen voi periaatteessa helposti toistaa ajamalla tutkijan käyttämän skriptin.



Kuva 1. AntConc-ohjelmalla on haettu CEEC-korpuksesta englannin *-ness*-johtimella muodostettuja sanoja eri kirjoitusasuineen (esim. *-nes*, *-nis*). Osumien joukossa on paljon roskaa, kuten erisnimiä ja *-ne*-loppuisten sanojen taivutusmuotoja.

Kun alustavat haut on tehty, on hakutulokset käytävä läpi väärien osumien poistamista ja tulosten luokittelua varten (ks. Laadullinen aineistopohjainen kielentutkimus tk.). Osa yllä mainituista korpuskäyttöliittymistä tukee tällaista luokittelua, mutta usein on helpompaa siirtää tulokset erilliseen tiedostoon, jonka voi sitten avata Excelissä tai muussa taulukkolaskentaohjelmassa ja tehdä annotoinnin siellä. Monista korpusohjelmista tällainen siirto- tai export-toiminto löytyykin, mutta esimerkiksi COHA-käyttöliittymästä tiedot on siirrettävä selaimen kopiointitoimintoa käyttäen. Hakutulosten luokittelu ja siivoaminen on usein tutkimuksen aikaa vievin osa.

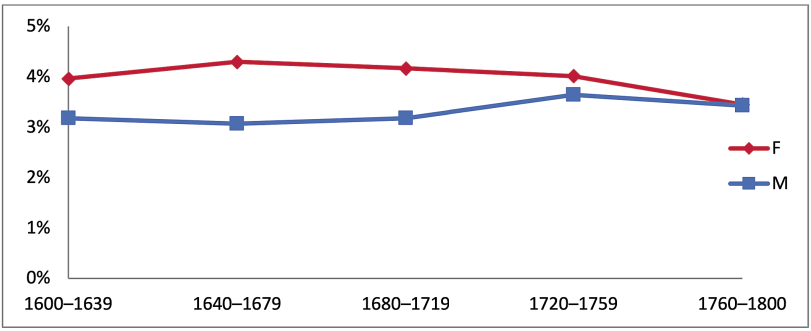
3.1.2. TILASTOLLINEN ANALYYSI JA VISUALISOINTI

Luokitelluista hakutuloksista voidaan siirtyä tekemään tilastollista analyysia. Tämä onnistuu valitusta menetelmästä riippuen joidenkin korpustyökalujen sisällä, taulukkolaskentaohjelmilla, erityisillä tilasto-ohjelmilla, kuten SPSS:llä[19], tai itse ohjelmoiden (viimeiseen vaihtoehtoon hyvä opaskirja on Gries 2013). Yleisin kvantitatiivinen menetelmä historiallisessa

korpuslingvistiikassa on laskea tutkittavan ilmiön **normalisoitu frekvenssi** aikakausittain ja vertailla frekvenssejä. Näin saadaan tietoa ilmiön esiintymistiheyden muutoksesta ajassa. Normalisoitu frekvenssi lasketaan jakamalla esiintymien lukumäärä koko aikakauden sanamäärällä; tämä voidaan kertoa vaikkapa kymmenellä tuhannella, jolloin saadaan ilmiön esiintyvyys kymmentä tuhatta sanaa kohti. Aikakausien sisällä voidaan tarkastella eri alakorpuksia, kuten genrejä tai sosiaalisia kategorioita, jolloin näiden normalisoitu frekvenssi lasketaan jakamalla alakorpuksen esiintymien lukumäärä alakorpuksen sanamäärällä. Esimerkiksi kuvassa 2 tarkastellaan englannin *I*-persoonapronominin normalisoidun frekvenssin muutosta ajassa naisten ja miesten alakorpuksissa. Kuvasta käy ilmi, että *I*-pronominin osuus on suurempi naisten kuin miesten kirjeissä kaikilla paitsi viimeisellä aikakaudella (ks. tietolaatikko 1).

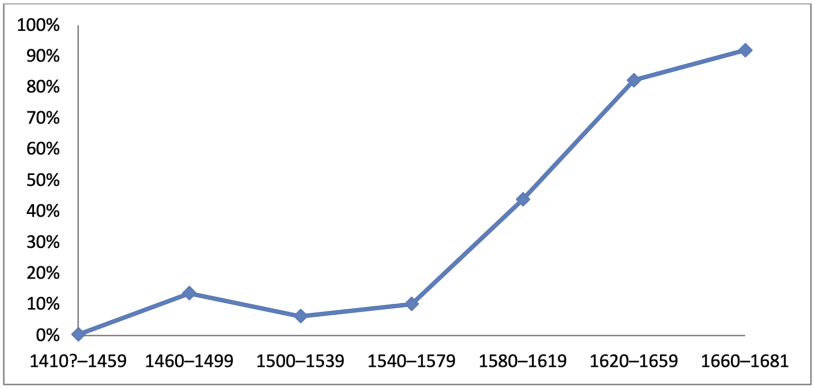
Tietolaatikko 1.

- Historiallinen sosiolingvistiikka tutkii historiallisen korpuslingvistiikan keinoin muun muassa sitä, miten kielenkäyttö vaihtelee eri yhteiskuntaryhmien kesken ja mitkä ryhmät johtavat kielenmuutosta (Nevalainen & Raumolin-Brunberg 2003).
- Vartiainen, Säily ja Hakala (2013) tutkivat englannin ensimmäisen ja toisen persoonan pronominien vaihtelua ja muutosta CEEC-korpuksessa. He havaitsivat, että naiset käyttivät näitä pronomineja keskimäärin enemmän kuin miehet. Tulos liittyy sukupuolittuneisiin kirjoitustyyleihin, joista on englannin kielestä evidenssiä vuosisatojen ajalta: miesten tyyli on usein informatiivisempi, naisten taas keskittyy enemmän vuorovaikutukseen.
- On kuitenkin huomattava, että tyylit eivät pysy täysin muuttumattomina: esimerkiksi vaikuttaa siltä, että 1700-luvulla naisten ja miesten kirjoitustyylit lähenivät toisiaan. Lisäksi erot tulevat esiin vain keskiarvoissa, ja yksilöiden välillä on paljon vaihtelua (ks. kuva 4). Lisäksi kielenkäyttöön vaikuttavat tarkemmat sosiaaliset roolit: esimerkiksi puolisoiden välisissä kirjeissä miehet saattoivat käyttää *I*-pronominia yhtä paljon kuin naiset (Vartiainen, Säily & Hakala 2013).



Kuva 2. Englannin *I*-persoonapronominin normalisoitu frekvenssi (ilmaistuna prosenttiosuutena kaikista sanoista) 1600-luvulta 1800-luvulle CEEC-korpuksessa. Aikakaudet on jaettu alakorpuksiin kirjoittajan sukupuolen perusteella (F = naiset, M = miehet).

Toinen normalisointimenetelmä otetaan käyttöön silloin, kun on olemassa vaihtoehtoisia tapoja sanoa sama asia. Esimerkiksi englannin yksikön 3. persoonan preesensin verbipäätteestä on ollut olemassa kaksi eri varianttia, *-s* ja *-th* (esim. *has* vs. *hath*). Nykyään vanhempaan *-th*-muotoon törmää lähinnä enää uskonnollisissa teksteissä ja mahdollisesti joissakin englannin murteissa. Variantit muodostavat kielellisen muuttujan eli variaabelin, *-s*/*-th*. Kun halutaan tutkia *-s*:n yleistymistä ajassa, lasketaan sen esiintymien osuus koko variaabelista eli *-s*:n ja *-th*:n yhteenlasketuista esiintymistä kullakin aikakaudella (kuva 3). Huomataan, että viivadiagrammi muistuttaa muodoltaan hieman S-kirjainta: aluksi muutos etenee hitaasti, sitten tulee nopea vaihe, ja lopuksi muutoksen eteneminen hidastuu jälleen. Monet muutokset noudattavat tällaista S-käyrää, mutta jotkin saattavat tehdä jopa U-käännöksen.



Kuva 3. Englannin yksikön 3. persoonan preesensin verbipääte *-s*:n osuus variaabelista *-s*/*-th* CEEC-korpuksessa, 1410?–1681 (Nevalainen & Raumolin-Brunberg 2003).

Korpuksen jakaminen aikakausiin eli **periodisaatio** on oma taiteenlajinsa, jossa tasapainotellaan kuvauksen tarkkuuden ja datan määrän välillä. Usein korpuksen kokoajat ovat alustavasti jakaneet korpuksen tietynmittaisiin aikakausiin, joille he myös ilmoittavat korpuksen manuaalissa sanamäärät. Kaudet voivat olla mielivaltaisen mittaisia, tai ne voivat perustua esim. kielen yleisesti tunnettuihin vaiheisiin, historiallisiin aikakausiin tai sukupolven keskimääräiseen pituuteen. Viimeinen vaihtoehto on yleinen sosiolingvistisissä korpuksissa, kuten CEEC:ssä, joskin aineiston vähyyden vuoksi sukupolviin perustuvat 20-vuotiskaudet joudutaan toisinaan yhdistämään 40-vuotiskausiksi (ks. kuvat 2 ja 3). Helsinki Corpus puolestaan on jaettu englannin yleisen kehityksen mukaisesti muinaisenglantiin (–1150), keskienglantiin (1150–1500) ja varhaisuusenglantiin (1500–1710). Nämä kaudet on jaettu edelleen noin sadan vuoden mittaisiin jaksoihin, jotka ovat sopivan kokoisia tutkittaviksi. Toisaalta esimerkiksi Gries ja Hilpert (2012) ovat esittäneet, että periodisaation tulisi olla muutoskohtaista ja perustua kunkin muutoksen omiin vaiheisiin; he käyttävät vaiheiden tunnistamiseen tilastollista klusterointimenetelmää (ks. Määrällinen korpuslingvistiikka tk.).

Jos normalisoiduissa frekvensseissä havaitaan muutosta ajassa, herää heti kysymys, onko kyseessä todellinen kielenmuutos vai voisiko

se johtua satunnaisesta vaihtelusta tai aineiston epätasaisuudesta. Normalisoitujen frekvenssien vertailun tukena voidaan käyttää **hypoteesitestausta** (esim. Coolidge 2013; ks. myös Määrällinen korpuslingvistiikka tk.), joka kertoo, onko kahden aikakauden välinen ero tilastollisesti merkitsevä. **Nollahypoteesina** on, että ero on sattumaa, ja ero on merkitsevä vain, mikäli todennäköisyys (*p*-arvo) sille, että nollahypoteesi hylätään väärin perustein, on valittua **merkitsevyystasoa** alhaisempi (tasona käytetään usein viittä prosenttia, $p < 0{,}05$). Todennäköisyys lasketaan **merkitsevyystestillä**, joita on useita erilaisia. Historiallisessa korpuslingvistiikassa on perinteisesti käytetty χ^2- eli khiin neliö -testiä, joka on helppo tehdä: tarvitsee tietää vain ilmiön esiintymien lukumäärät kummaltakin aikakaudelta sekä aikakausien kokonaissanamäärät, ja *p*-arvon voi katsoa taulukosta tai käyttää nettilaskuria.

Valitettavasti χ^2-testi ja muut niin kutsutut sanasäkkitestit (engl. *bag-of-words tests*) eivät ota huomioon sanojen jakautumista korpuksessa, vaan ne olettavat, että sanat esiintyvät toisistaan riippumatta, mikä ei tietenkään pidä paikkaansa. Tämä tekee testeistä ylioptimistisia, eli ne pitävät eroa usein virheellisesti merkitsevänä. Oletetaan esimerkiksi, että korpuksessa on kaksi aikakautta, joissa kummassakin on 50 000 juoksevaa sanaa. Jos jokin sana esiintyy ensimmäisellä aikakaudella vaikkapa 25 kertaa ja toisella aikakaudella vain 5 kertaa, tällaiset testit pitävät eroa käytännössä aina merkitsevänä (taulukko 1). Testit eivät kuitenkaan ota huomioon sitä mahdollisuutta, että ensimmäisen aikakauden 25 esiintymää voivat olla kaikki samassa tekstissä, jolloin korkea frekvenssi on pelkkää sattumaa (Säily 2014, 49).

	Aikakausi 1	**Aikakausi 2**
Sana *X*	25	5
Jokin muu sana kuin *X*	49 975	49 995

Taulukko 1. Esimerkki khiin neliö -testissä käytettävästä kontingenssitaulusta, johon laitetaan sanamäärätiedot. Tilasto-ohjelman antama *p*-arvo on tässä tapauksessa 0,0005, eli χ^2-testin mukaan on vain 0,05 %:n mahdollisuus, että nollahypoteesi hylätään väärin perustein.

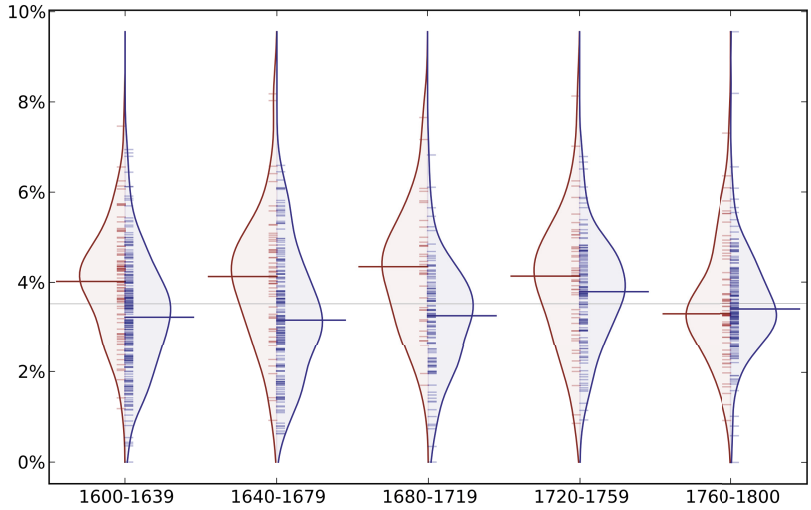
Näistä syistä korpuslingvistiikassa ollaankin siirtymässä ns. hajontatietoisiin testeihin (engl. *dispersion-aware tests*), joihin lukeutuvat mm. t-testi, Wilcoxonin järjestyssummatesti, Kendallin tau -testi ja uudelleenotantaan perustuvat testit (Säily 2014, 46). Näitä testejä pystytään käyttämään siten, että ne huomioivat sanojen jakautumisen korpuksessa, mutta siinä tapauksessa niiden käyttö on hankalampaa, sillä syötteeksi eivät riitä aikakausitason frekvenssit vaan on mentävä tekstitasolle: pitää tietää ilmiön frekvenssi kussakin korpuksen tekstissä sekä kunkin tekstin kokonaissanamäärä. Testeistä aloittelijalle helpoin ja luotettavin lienee Wilcoxonin järjestyssummatesti, toiselta nimeltään Mann-Whitneyn U-testi (esim. Metsämuuronen 2011, 386–390, 1102–1107), jonka pystyy tekemään Excelissä ja jolle on myös nettilaskuri. Tilastollisen merkitsevyyden osoittamiseen liittyy toki myös monimutkaistavia tekijöitä: mitä useampia hypoteeseja testataan, sitä suuremmalla todennäköisyydellä nollahypoteesi kumotaan väärin perustein. Tämän välttämiseksi voidaan käyttää erilaisia menetelmiä, jotka käytännössä tiukentavat merkitsevyystasoa (esim. Säily 2014, 49–51).

Historiallisessa korpuslingvistiikassa voidaan käyttää myös sofistikoituneempia tilastollisia menetelmiä, joiden avulla pystytään analysoimaan useiden ilmiöiden muutosta tai useita yhden ilmiön esiintyvyyteen vaikuttavia tekijöitä yhtä aikaa. Tällaista **monimuuttuja-analyysia** on käytetty esimerkiksi rakennevaihtelun, kuten englannin *-s-* ja *of*-genetiivien vaihtelun, tutkimiseen (Szmrecsanyi, Rosenbach, Bresnan & Wolk 2014) sekä genremuutoksen tutkimiseen Douglas Biberin kehittämää piirrepatteristoa käyttäen (Biber 1988; Biber & Finegan 1997). Biberin multidimensionaalinen menetelmä laskee kymmenien kielenpiirteiden frekvenssit teksteissä ja etsii niistä faktorianalyysin keinoin piirteet, joiden frekvenssit vaihtelevat samalla tavoin tekstien välillä. Näillä piirteillä tulkitaan olevan jokin kommunikatiivinen syy esiintyä yhdessä, ja eri aikakausia ja genrejä voidaan vertailla sijoittamalla ne piirteiden muodostamille rekisteriulottuvuuksille. Tämäntyyppinen tutkimus osoittaa, että kaikki muutos aineistossa ei ole kielenmuutosta, vaan se voi liittyä myös genrekonventioiden muutokseen. Esimerkiksi englannin historiassa on ollut käynnissä jo vuosisatoja kestänyt prosessi, jossa genret epämuodollistuvat (engl. *colloquialization*), jolloin niissä käytetään yhä

puhekielenomaisempia piirteitä, kuten persoonapronomineja sekä ihmisen ajatuksiin ja mielipiteisiin liittyviä verbejä (esim. *think*, *guess*).

Nykypäivän korpuslingvistiikassa yleiset avainsana-analyysi ja kollokaatioiden tutkimus (ks. Määrällinen korpuslingvistiikka tk.) ovat historiallisessa korpuslingvistiikassa olleet harvinaisempia, osittain aineiston vähäisyyden, osittain taas sanojen kirjoitusasujen vaihtelun takia. Ennen oikeinkirjoituksen standardisointia sanojen kirjoitusasuissa oli valtava määrä vaihtelua – esimerkiksi Parsed Corpus of Early English Correspondence[20] -kirjekorpuksessa (1410–1681) englannin kielen *tomorrow*-sanaa on kirjoitettu ainakin seuraavilla tavoilla: *to marrow*, *to moroe*, *to moroughe*, *to morow*, *to morowe*, *to morroughe*, *to morrow*, *to morrowe*, *to morue*, *to morwe*, *to-morow*, *to-morowe*, *to-morrow*, *to-morrowe*, *to-morw*, *to-morwe*, *tomorow*, *tomorowe*, *tomorrow*, *tomorrowe*, *too morrow*, *toomorrow*. Joistakin historiallisista korpuksista on nyttemmin tehty normalisoituja versioita, joissa tekstien kirjoitusasu on standardisoitu esim. VARD-ohjelmalla (Baron 2011); automaattinen standardisointi ei kuitenkaan pure harvinaisempiin kirjoitusasuihin, joten osa korpuksien sanoista jää edelleen standardisoimatta. Yksi harvoista historiallisista avainsana-analyyseista on Lijffijt, Säily & Nevalainen (2012); kollokaatioiden muutosta puolestaan ovat tutkineet mm. Baker, Brezina ja McEnery (2017), jotka keskittyvät 1800–1900-lukuihin, jolloin englannin kirjoittaminen oli etenkin painetuissa teksteissä jo melko yhtenäistä.

Tilastollisten menetelmien rinnalle tärkeiksi ovat nousemassa **tiedon visualisoinnin** menetelmät. Visualisointi on tiedon esittämistä tiivistetyssä muodossa; on osoitettu, että ihminen saa tietoa näköaistin kautta enemmän kuin kaikkien muiden aistien kautta yhteensä (Ware 2004). Yksinkertaisia, kuvien 2 ja 3 kaltaisia viivadiagrammeja voi hyvin tehdä vaikkapa Excelissä, mutta vaativampia visualisointeja varten tarvitaan jotain muuta, kuten R-ympäristössä toimiva ggplot2-kuvantamisjärjestelmä.[21] Historiallisessa korpuslingvistiikassa on nähtävissä suuntaus viivadiagrammeista periodin sisäistä vaihtelua ilmentäviin kuviin, joista yhtenä esimerkkinä mainittakoon *beanplot* eli papukaavio (kuva 4; Kampstra 2008).

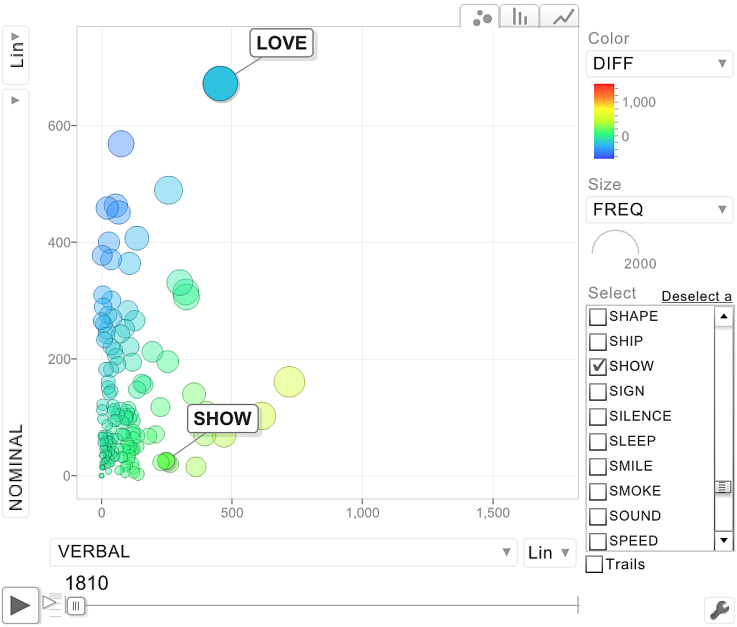


Kuva 4. Englannin I-persoonapronominin vaihtelu ja muutos CEEC-korpuksessa esitettynä *beanplot*-kaaviona (vrt. kuva 2). Naiset punaisella, miehet sinisellä. (Vartiainen, Säily & Hakala 2013.)

Kuva 4 pohjautuu samaan aineistoon kuin kuva 2, mutta se antaa paljon enemmän informaatiota. Paksut vaakaviivat kertovat *I*-pronominin mediaanifrekvenssin naisilla (vas.) ja miehillä (oik.) aikakausittain, kun taas kukin pienempi vaakaviiva ilmaisee pronominin normalisoidun frekvenssin yksittäisen henkilön kirjoittamissa teksteissä (mediaani on näistä havainnoista keskimmäinen). Näemme siis yhdellä silmäyksellä muutoksen yleiskuvan lisäksi sen, moneenko eri havaintoon mediaanit perustuvat: esimerkiksi keskimmäisellä kaudella naisten alakorpuksessa on suhteellisen vähän havaintoja, joten emme välttämättä luota sen mediaaniin yhtä paljon kuin vaikkapa ensimmäisen kauden. Wilcoxonin järjestyssummatestin mukaan naisten ja miesten välinen ero on tilastollisesti merkitsevä kahdella ensimmäisellä kaudella (Vartiainen, Säily & Hakala 2013, 240). Kaaviosta näkyy myös, miten havainnot jakautuvat alakorpuksien sisällä. Kunkin alakorpuksen havaintojen ympärille on piirretty jakauma (joka kaavion kehittäjän mielestä muistuttaa pavun palkoa), joka osoittaa, mihin suurin osa havainnoista keskittyy. Tämä auttaa huomaamaan aineiston epänormaaliudet: jos esimerkiksi jokin

osajoukko havainnoista käyttäisi pronominia huomattavan paljon ja jokin toinen taas vähän, jakaumassa näkyisi kaksi erillistä keskittymää.

Uusinta uutta korpuslingvistiikassa on interaktiivinen visualisointi, ja historiallinen korpuslingvistiikka on tässäkin ajan hermolla. Esimerkiksi Hilpert (2011) käyttää animoituja pistekaavioita (engl. *motion charts*), joissa kielenmuutos näkyy eri kielen ilmiöitä esittävien pisteiden liikkeenä kaaviossa. Pisteen koko kertoo sen normalisoidun frekvenssin. Kuvassa 5 näkyy joukko englannin sanoja, joita voidaan käyttää sekä substantiiveina (engl. *nominal*, y-akseli) että verbeinä (engl. *verbal*, x-akseli). Kuvasta ilmenee vuoden 1810 tilanne COHA-korpuksessa, mutta vasemman alakulman nuolta painamalla alkaa animaatio, joka osoittaa, miten sanojen käyttö muuttuu ajassa aina 2000-luvulle asti. Esimerkiksi *love*-sanaa käytetään aluksi lähinnä substantiivina, mutta



Kuva 5. Animoitu pistekaavio ambikategoristen sanojen muutoksesta COHA-korpuksessa (http://members.unine.ch/martin.hilpert/motion.html).

ajan myötä se muuttuu yhä ambikategorisemmaksi, eli sitä käytetään yhtä lailla sekä substantiivina (*What is love?*) että verbinä (*I love you*). Animoidut pistekaaviot auttavat tutkijaa prosessoimaan paljon tietoa tehokkaasti, ja niiden avulla on helppo saada yleiskuva muutoksesta sekä toisaalta seurata yksittäisten sanojen tai ilmiöiden muutosta.

On huomattava, että numerot tai kuvat itsessään eivät ole vielä tutkimustulos, vaan niiden tulkinnassa tarvitaan myös laadullista tutkimusta. Esimerkiksi pronominitutkimuksessamme (kuvat 2 ja 4, tietolaatikko 1) palasimme toistuvasti korpusteksteihin: luimme miesten ja naisten kirjoittamia kirjeitä saadaksemme selville, mistä sukupuoliero persoonapronominien käytössä johtui. Tarkastelimme, miten miehet ja naiset käyttivät pronomineja ja miten tämä liittyi heidän sosiaalisiin rooleihinsa isinä, poikina, aviomiehinä, vaimoina ja äiteinä. Havaitsimme muun muassa, että vaikka 1600-luvun miehet ja naiset käyttivät puolisoiden välisissä kirjeissä *I*-pronominia keskimäärin yhtä paljon, he tekivät pronominilla eri asioita: aviomiehet kertoivat tyypillisesti siitä, mitä he olivat tehneet (esim. *I met my old friend*), kun taas vaimot kertoivat omista ajatuksistaan ja tunteistaan (esim. *I think*); toisaalta vaimot käyttivät enemmän *you*-pronominia, koska hekin keskittyivät aviomiehen tekemisiin. Lisäksi kytkimme löydökset historiallisiin konteksteihinsa, kuten siihen, että aviopuolisoiden välinen suhde oli tuon ajan Britanniassa epätasa-arvoinen ja mies oli suhteen hallitseva osapuoli. Seuraava luku tarjoaa lisää esimerkkejä siitä, mitä historiallisen korpuslingvistiikan keinoin on tutkittu.

3.2. Mitä metodilla on tutkittu

3.2.1. KIELIOPILLISTUMINEN

Kieliopillistuminen on ollut yksi keskeisimmistä historiallisen korpuslingvistiikan menetelmin tutkituista kielenmuutoksista. *Kieliopillistumisella* tarkoitetaan yleisesti ottaen ilmiötä, jossa olemassa olevan leksikaalisen sanan tai konstruktion merkitys muuttuu asteittain kieliopillisemmaksi: konstruktiot, joilla on aiemmin viitattu konkreettisiin asioihin tai tapahtumiin, valjastetaan kuvaamaan yhä abstraktimpia

merkityksiä, kuten aikamuotoja, aspektia tai lukua. Koska tällaiset kieli-oppikategoriat ovat pakollinen osa kielen järjestelmää siinä mielessä, että kielenkäyttäjän on tyypillisesti aina kytkettävä puhunnoksensa esimerkiksi oman kielensä tempus- tai aspektijärjestelmään valitsemalla tarkoitukseensa sopiva muoto, kieliopillistuneen konstruktion käyttötaajuus on valtavasti suurempi kuin mitä sillä oli ennen kieliopillistumista. Yleistyneen käytön seurauksena konstruktiossa tapahtuu usein fonologisia ja morfologisia muutoksia, joissa se muuttuu foneettisesti yksinkertaisemmaksi sekä mahdollisesti myös vapaasta morfeemista sidonnaiseksi morfeemiksi.

Esimerkiksi voidaan ottaa vaikkapa muinaisenglannin verbi *willan*. Se oli alkujaan leksikaalinen verbi, joka merkitsi 'tahtoa' tai 'haluta', kuten esimerkissä 1.

1.

ða	*cwæþ*	*he*	*to*	*him,*	***wylt***	*þu*	*hal*	*beon?*
sitten	sanoa. 3SG.PRET	hän	PREP	hän. DAT.M.SG	haluta. 2SG.PRES	sinä	terve	olla/ tulla. INF[22]

'Sitten hän sanoi hänelle: haluatko tulla terveeksi?' (HC, 950–1050)

Vuosisatojen saatossa *willan*-verbiä alettiin käyttää yleisemmin ilmaisemaan tulevia, ennustettavia tapahtumia ja lopulta pelkästään tulevaa aikaa. Prosessin lopputuloksena verbistä tuli futuurisuutta ilmaiseva apuverbi (nykyenglannissa *will*), joka varsinkin puheessa esiintyy usein muodossa *'ll* (vrt. *we will*; *we'll*).

Toinen esimerkki kieliopillistumiselle ominaisista semanttisista ja rakenteellisista muutoksista ovat englannin kielen binominaaliset kvanttorit (esim. *a lot of* N, *a bit of* N). Sekä *lot* että *bit* ovat näissä rakenteissa viitanneet alkujaan konkreettisiin asioihin: *lot* maapalstaan (esim. *a lot of land*) ja *bit* haukkapalaan (esim. *a bit(e) of bread*). Kieliopillistumisprosessissa kielenpuhujat tulkitsivat näiden sanojen merkityksen uudelleen, niin että niiden tarkoitteeseen implisiittisesti liittynyt pieni/suuri

koko tai määrä nousikin keskeiseksi, ja tästä seurasi konstruktioiden rakenteellinen uudelleentulkinta: *lot-* ja *bit(e)*-sanat eivät enää olleetkaan substantiivilausekkeidensa pääsanoja, vaan ne käsitettiin syntaktisesti osaksi monisanaista kvanttorirakennetta. Tällöin alun perin jälkimääritteen osana olleesta substantiivista tuli uudelleen tulkitun lausekkeen pääsana: [a lot] [of land] → [a lot of] [land]. Kieliopillistumista voidaan tarkastella myös käyttökontekstien laajentumisen näkökulmasta. Esimerkiksi *a lot (of)-* ja *a bit (of)* -kvanttoreita on alettu käyttää myös muunlaisten kieliopillisten merkitysten ilmaisuun, kuten määrää kuvaavina adverbeina (esim. *I like you a lot*) ja asteadverbeina (esim. *a bit rude*). Samoin kuin *will*-apuverbistä myös näistä konstruktioista on olemassa kontrahoidut muodot: varsinkin *lotta* on yleinen, mutta myös *bitta*-muotoa käytetään etenkin puheessa. Sama koskee myös monia muita englannin binominaalisia kvanttoreita, kuten *sort of* (*sorta*), *kind of* (*kinda*) ja *type of* (*typa*).

Yksi tutkituimmista kieliopillistumisprosesseista on englannin *be going to* -konstruktion muuttuminen liikettä kuvaavasta rakenteesta futuuria ilmaisevaksi konstruktioksi. Esimerkit 2 ja 3 kuvaavat rakenteen käyttöä varhaisuusenglannissa juuri ennen kieliopillistumisprosessin alkua, kun taas 4 ja 5 ovat esimerkkejä nykyenglannissa yleisestä, kieliopillistuneesta futuurista. Esimerkki 5 on kiistaton todiste rakenteen kieliopillistumisesta: leksikaalinen *go*-verbi on lauseessa toistettu, koska *be going to* -rakenteessa olevaa *going*-sanaa ei enää käsitetä liikettä ilmaisevaksi verbiksi.

2. *A yong mayd* ***going to*** *a feast with hir mother…* (PPCEME[23], 1582)
3. *Mistris Page, trust me, I* ***was going to*** *your house.* (PPCEME, 1599)
4. *It's* ***going to*** *be tough for them.* (COCA[24], 2012)
5. *I'****m going to*** *go to Nashville, Tennessee.* (COCA, 2012)

Esimerkit 2–5 ilmentävät *be going to* -konstruktion muutoksen alku- ja päätepisteitä, mutta rakenteen kieliopillistuminen tapahtui tosiasiassa vaiheittain vuosisatojen saatossa. Muutos sai alkunsa, kun kielenpuhujat alkoivat 1600-luvulla korostaa fyysiseen liikkeeseen tyypillisesti liittyvää tahdonalaisuutta ja aikomusta siinä määrin, että konstruktion ei

enää välttämättä tarvinnutkaan kuvata fyysistä liikettä. Esimerkeissä 6 ja 7 on toki mahdollista, että puhuja liikkuu paikasta toiseen puhumaan, mutta todennäköisempää on, että hän ilmaisee ainoastaan aikomuksensa puhua.

6. *now, a certaine religious Woeman dyed at home with vs, about a day and a halfe, before that occurred, whereof i* ***am going to*** *speake* (EEBO, 1642)
7. *but there is another thing worth your observation, which i* ***am going to*** *tell you* (EEBO, 1653)

Leuvenin yliopiston tutkijat Sara Budts ja Peter Petré tarkastelivat vuonna 2016 julkaistussa korpustutkimuksessaan *be going to* -konstruktion kieliopillistumista myöhäisuusenglannin kaudella (1700–1920). Myös he tulkitsivat *be going to* -rakenteen kieliopillistumisen subjektifikaatioprosessiksi, jossa konstruktion alkuperäinen, liikettä kuvaava objektiivinen merkitys muuttui ensin subjektiivisemmaksi (puhujan omaa intentiota tarkoittavaksi) ja myöhemmin vielä subjektiivisemmaksi, niin että puhujat alkoivat käyttää konstruktiota ennakoidessaan yleisemmin tulevia tapahtumia ja asioita tai ollessaan epävarmoja tulevista tapahtumista. Budts ja Petré havaitsivat muun muassa, että ensimmäisen persoonan subjektipronominien osuus kaikista *be going to* -rakenteiden subjekteista pieneni tutkitun kauden kuluessa, kun taas toisen persoonan subjektien osuus kasvoi. Tämä selittyy muun muassa kysymyslauseiden yleistymisellä: toiselle ihmiselle esitetyt kysymykset, kuten *are you going to come tomorrow?*, ovat esimerkki rakenteen uudenlaisesta käytöstä, jossa puhuja ei ilmaise omaa intentiotaan vaan osoittaa epävarmuutensa toisen henkilön aikomuksista. Budts ja Petré havaitsivat aineistossaan myös jyrkän laskun tulevan tapahtuman pikaisuudessa: vuosina 1710–1745 peräti 79 % *be going to* -rakenteista viittasi puhetilannetta välittömästi seuraavaan tapahtumaan, kun taas 1886–1920 vastaava osuus oli enää 29 %. Tilastollisena menetelmänä Budts ja Petré käyttivät Kendallin tau -testiä.

Tietolaatikko 2.

- Budtsin ja Petrén käyttämät korpukset olivat Penn Parsed Corpus of Modern British English (Kroch, Santorini & Diertani 2016), ECCO-TCP Corpus[25] ja Corpus of Late Modern English Texts v.3.0 (Diller, De Smet & Tyrkkö 2011).
- Useamman kuin yhden korpuksen käyttäminen historiallisessa korpustutkimuksessa on varsin yleistä; yksittäinen korpus voi olla liian suppea tarkasteltavan kielen ilmiön kattavaan tutkimiseen.
- Tutkijan on otettava tarkoin huomioon tutkimuksessaan käyttämiensä korpusten väliset erot. Koska suurin osa korpuksista on koottu eri periaatteita noudattaen, ei esimerkiksi sanojen frekvenssien vertailu korpusten välillä ole ongelmatonta.

Budtsin ja Petrén tutkimus on monessa mielessä tyyppiesimerkki historiallisen korpuslingvistiikan menetelmin toteutettavasta tutkimuksesta: tutkijat pureutuvat tarkastelemaansa ilmiöön analysoimalla kattavasti sen eri käyttökonteksteja ja niissä tapahtuvia mikrotason muutoksia, minkä ansiosta he pystyvät tarjoamaan yksityiskohtaisen analyysin kieliopillistumisprosessiin liittyvistä semanttisista tekijöistä.

3.2.2. Historiallinen sosiolingvistiikka: *-s* ja *-th* verbin kolmannen persoonan päätteinä englannin kielessä

Historiallinen sosiolingvistiikka on kielitieteen suuntaus, joka sai alkusysäyksensä sosiolingvististen hypoteesien testaamisesta historiallisissa aineistoissa. Näitä hypoteeseja olivat esimerkiksi naisten johtava rooli kielenmuutoksessa sekä ajatus siitä, että vertailemalla eri-ikäisten ihmisten kielenkäyttöä saadaan samalla tietoa kielenmuutoksesta ajassa: ihmisten kielenkäytön oletetaan suurelta osin vakiintuvan jo nuorena, joten vanhempien ihmisten kielenkäyttö edustaa varhaisempaa aikakautta kuin nuorempien (ns. näennäisaikahypoteesi). Suzanne

Romainen (1982) ohella historiallisen sosiolingvistiikan pioneereja ovat olleet Helsingin yliopiston tutkijat Terttu Nevalainen ja Helena Raumolin-Brunberg, joiden johdolla myös koottiin CEEC-korpus.

Nevalainen ja Raumolin-Brunberg (2003) tutkivat useisiin eri kielellisiin muuttujiin eli variaabeleihin pohjautuvia muutoksia CEEC-korpuksessa aikavälillä 1410–1681 historiallisen sosiolingvistiikan näkökulmasta. Otamme tähän esimerkiksi kuvassa 3 esitellyn englannin yksikön kolmannen persoonan preesensin verbipäätteen variaabelin *-s*/*-th*. Nevalainen ja Raumolin-Brunberg huomauttavat, että kyseessä on oikeastaan kaksi eri muutosta: toisaalta painottoman /e/-vokaalin kato päätteestä, jonka alkuperäinen muoto oli siis *-eth*, ja toisaalta viimeisen konsonantin muuttuminen sibilantiksi. Nevalainen ja Raumolin-Brunberg osoittavat, että molemmat muutokset alkoivat Pohjois-Englannista. Ensimmäisenä alkoi *-eth*:n muuttuminen *-es*:ksi (esim. *eateth* vs. *eates*, joskin vokaalia saatettiin kirjoittaa esimerkiksi *y*:llä), mikä näkyy kuvassa 3 pienenä nousuna 1400-luvun lopulla; tämä ei kuitenkaan kunnolla levinnyt etelämmäksi. Noin sata vuotta myöhemmin *-s*-variantti aloitti uuden nousun, joka pohjautui *-eth*-muodon ja vokaalittoman *-s*-muodon väliseen vaihteluun (esim. *eateth* vs. *eats*); tuon ajan kirjoittajat pitivätkin *-s*-muotoa *-eth*-muodon lyhenteenä. Tällä kertaa muutos eteni useimmissa murteissa loppuun asti, ja nykyenglannin puhujat tuntevat *-th*-variantin lähinnä Shakespearen näytelmistä ja Jaakko-kuninkaan aikaisesta raamatunkäännöksestä (1611), joka on edelleen käytössä joissakin kirkkokunnissa.

Nevalainen ja Raumolin-Brunberg (2003) tarkastelivat muutosta useiden sosiaalisten kategorioiden suhteen. Kirjoittajan asuinpaikan lisäksi variaabelin käyttöön – eli siihen, kumpaa varianttia kirjoittaja käytti enemmän – saattoivat vaikuttaa mm. hänen sukupuolensa, yhteiskunnallinen asemansa sekä suhteensa kirjeen vastaanottajaan. Menetelmänä Nevalainen ja Raumolin-Brunberg käyttivät variantin normalisoidun frekvenssin laskemista, kuten kuvassa 3, mutta he jakoivat aikakaudet sosiaalisten kategorioiden perusteella muodostettuihin alakorpuksiin. Sitten he vertailivat alakorpuksien normalisoituja frekvenssejä hypoteesitestauksen keinoin χ^2-testillä. Yllä mainitun alueellisen vaihtelun lisäksi he havaitsivat, että muutosta johtivat *-s*:n uuden nousun osalta odotetusti naiset.

Nopeita *-s*:n omaksujia olivat etenkin säätyläisiä alhaisemmassa yhteiskunnallisessa asemassa olleet kirjoittajat sekä toisaalta sosiaaliset nousijat. Täydentääkseen tuloksiaan Nevalainen ja Raumolin-Brunberg toteuttivat monimuuttuja-analyysin VARBRUL-ohjelmalla, jonka eri versioita on käytetty yleisesti sosiolingvistisessä tutkimuksessa.[26] Monimuuttuja-analyysin avulla pystytään arvioimaan eri sosiaalisten kategorioiden suhteellista tärkeyttä muutoksessa. Tässä analyysissa kirjoittajan asuinalue ja sukupuoli nousivat tärkeiksi muutosta selittäviksi tekijöiksi, kun taas kirjoittajan suhteella kirjeen vastaanottajaan (perheenjäsen/ystävä vs. muu) ei havaittu olevan merkittävää vaikutusta.

4. Yhteenveto

Historiallinen korpuslingvistiikka tarjoaa tutkijalle monipuolisen menetelmän kielen varhaisempien vaiheiden ja kielenmuutoksen systemaattiseen tarkasteluun. Korpuslingvistin työssä korostuu tutkimuksen lähteinä käytettyjen korpusten tarkka tunteminen, mutta kuten olemme pyrkineet tuomaan tässä artikkelissa esille, aineiston tunteminen on vain ennakkoedellytys tutkimuksen tekemiselle. Kieli voi muuttua monesta syystä, ja historiallisen korpuslingvistin työkalupakkiin on hyvä kuulua esimerkiksi sosiolingvistiikan, pragmatiikan, kielikontaktien sekä kognitiivisen kielitieteen teorioiden hallinta. Yleisesti ottaen historialliselle korpuslingvistiikalle leimaa-antavaa on määrällisten ja laadullisten menetelmien yhdistäminen: tutkijan korpusaineistosta laskemien frekvenssien merkitsevyyttä voidaan esimerkiksi tarkastella tilastollisin menetelmin, ja niitä voidaan havainnollistaa monenlaisten visualisointien avulla, mutta tulosten tulkinta nojaa aina laadullisen tutkimuksen periaatteisiin – numerot itsessään eivät ole vielä tutkimustulos. Toisaalta on hyvä pitää mielessä, että korkealaatuista korpuslingvistiikkaa on mahdollista tehdä myös ilman sofistikoituneita tilastollisia menetelmiä: kielenmuutokset voivat olla niin selkeitä, että niitä on mahdollista seurata pelkkien frekvenssitietojen avulla.

Tässä artikkelissa annetut esimerkit historiallisen korpuslingvistiikan kentästä on tarkoitettu ensikatsaukseksi metodologian eri soveltamismahdollisuuksiin. On selvää, että menetelmää käytetään laajalti myös muunlaisten tutkimuskysymysten tarkasteluun: esimerkiksi historiallista semantiikkaa, pragmatiikkaa, fonologiaa ja kielen murrevaihtelua on kaikkia tutkittu monipuolisesti korpuslingvistiikan menetelmin. **Digitaalisten ihmistieteiden** yleistymisen myötä on todennäköistä, että niin korpuksia kuin muitakin digitaalisia kieliaineistoja on tulevaisuudessa tutkijan käytössä entistä enemmän ja nämä aineistot ovat yhä useammin saatavilla myös muussa kuin tekstimuodossa. Esimerkiksi vanhojen kirjeiden digitoidut kuvat voivat tarjota paljon sellaista tietoa tekstistä ja sen tuottamiseen liittyvistä olosuhteista, joita pelkkä tietokoneen ruudulla näkyvä tekstirivi ei tavoita. Kuten CEEC-korpuksen esimerkki osoittaa, myös tekstien kirjoittajiin liittyviä taustatietoja on mahdollista kytkeä korpusaineistoon, jolloin ne ovat hyödyllisiä paitsi kielentutkijoille myös historioitsijoille. Muitakin tällaisia aineistoja on jo olemassa (esim. London Lives[27]), joskin ne on tyypillisesti koottu pääasiassa historiantutkimuksen tarpeisiin, eikä niiden käyttö sellaisenaan ole täysin ongelmatonta korpuslingvistiikan näkökulmasta.

Tämän lyhyen katsauksen pyrkimyksenä on ollut tarjota joitakin näkökulmia historiallisen korpuslingvistiikan taustaan, perusolettamuksiin ja viimeaikaisiin kehityskulkuihin. Luvussa 3 esitetyt tapaustutkimukset ovat esimerkkejä yksittäisiin kysymyksiin pureutuvasta tutkimuksesta, jotka osana suurempaa kokonaisuutta voivat tarjota uutta tietoa esimerkiksi sellaisista perustavanlaatuisista kysymyksistä, joita nostimme esiin artikkelimme johdannossa. Toivomme, että tämä artikkeli välittää lukijalle kuvan korpuslingvistiikasta monipuolisena menetelmänä, joka kehittyy niin kielitieteen tutkimusalan kuin yleisen tietoteknisen kehityksen rinnalla ja jonka avulla voimme saada jatkuvasti uutta tietoa paitsi kielestä järjestelmänä myös ihmisyhteisöjen kielikäytäntöjen vaihtelusta ja muutoksesta.

Aiheesta lisää:

Hopper, Paul J. & Traugott, Elizabeth Closs. 2003. *Grammaticalization*. (2. p.). Cambridge: Cambridge University Press.

Krug, Manfred & Schlüter, Julia (toim.). 2013. *Research Methods in Language Variation and Change*. Cambridge: Cambridge University Press.

McEnery, Tony & Hardie, Andrew. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press.

Nevalainen, Terttu & Raumolin-Brunberg, Helena. 2003. *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. Longman Linguistics Library. London: Pearson Education.

VIITTEET

1 http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/.
2 http://www.helsinki.fi/varieng/CoRD/corpora/FROWN/.
3 http://www.helsinki.fi/varieng/CoRD/corpora/LOB/.
4 http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/.
5 http://www.helsinki.fi/varieng/CoRD/corpora/BLOB-1931/.
6 http://www.helsinki.fi/varieng/CoRD/corpora/B-BROWN/.
7 http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/.
8 https://www.english-corpora.org/coha/.
9 https://www.english-corpora.org/eebo/, https://digi.kansalliskirjasto.fi.
10 http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/.
11 http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/, http://kaino.kotus.fi/korpus/vks/meta/vks_coll_rdf.xml.
12 https://www.kielipankki.fi.
13 http://www.meta-share.org, https://www.clarin.eu/content/language-resource-inventory, http://www.helsinki.fi/varieng/CoRD/.
14 https://www.corpusdelespanol.org, https://www.corpusdoportugues.org.
15 https://korp.csc.fi.
16 https://cqpweb.lancs.ac.uk/.
17 http://www.lexically.net/wordsmith/, http://www.laurenceanthony.net/software/antconc/.
18 https://www.r-project.org.
19 https://www.ibm.com/analytics/spss-statistics-software.

20 http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/.
21 https://ggplot2.tidyverse.org.
22 Lyhenteet: 3SG = yksikön 3. persoona, PRET = preteriti, PREP = prepositio, DAT = datiivi, M = maskuliini, 2SG = yksikön 2. persoona, PRES = preesens, INF = infinitiivi.
23 PPCEME = Penn-Helsinki Parsed Corpus of Early Modern English (https://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-3/).
24 COCA = Corpus of Contemporary American English (https://www.english-corpora.org/coca/).
25 https://textcreationpartnership.org/tcp-texts/ecco-tcp-eighteenth-century-collections-online/.
26 https://en.wikipedia.org/wiki/Variable_rules_analysis.
27 https://www.londonlives.org.

KIRJALLISUUS

Baker, Helen, Brezina, Vaclav & McEnery, Tony. 2017. Ireland in British Parliamentary debates 1803–2005: Plotting changes in discourse in a large volume of time-series corpus data. Julkaisussa: Säily, Tanja, Nurmi, Arja, Palander-Collin, Minna & Auer, Anita (toim.) *Exploring Future Paths for Historical Sociolinguistics*. Advances in Historical Sociolinguistics 7. Amsterdam: John Benjamins, 83–107.

Baron, Alistair. 2011. VARD 2. Tietokoneohjelma. Saatavissa: http://ucrel.lancs.ac.uk/vard/.

Benveniste, Emile. 1971 [1958]. Subjectivity in language. Julkaisussa: *Problems in General Linguistics*. Englanninkielinen käännös Mary Elizabeth Meek. Coral Gables (FL): University of Miami Press, 223–230.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas & Finegan, Edward. 1997. Diachronic relations among speech-based and written registers in English. Julkaisussa: Nevalainen, Terttu & Kahlas-Tarkka, Leena (toim.) *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. Mémoires de la Société Néophilologique de Helsinki 52. Helsinki: Société Néophilologique, 253–275.

Budts, Sara & Petré, Peter. 2016. Reading the intentions of *be going to*: On the subjectification of future markers. *Folia Linguistica Historica* 37:1, 1–31.

Coolidge, Frederick L. 2013. *Statistics: A Gentle Introduction*. (3. p.). Thousand Oaks (CA): Sage.

Diller, Hans-Jürgen, De Smet, Hendrik & Tyrkkö, Jukka. 2011. A European database of descriptors of English electronic texts. *The European English Messenger* 19, 21–35.

Francis, W. Nelson & Kučera, Henry. 1979. *Manual to Accompany a Standard Sample of Present-day Edited American English, for Use with Digital Computers*. Alkuperäinen p. 1964, korjattu 1971, korjattu ja täydennetty 1979. Providence, Rhode Island: Department of Linguistics, Brown University.

Gries, Stefan Th. 2013. *Statistics for Linguistics with R: A Practical Introduction*. (2. p.). Berlin: De Gruyter Mouton.

Gries, Stefan Th. & Hilpert, Martin. 2012. Variability-based Neighbor Clustering: A bottom-up approach to periodization in historical linguistics. Julkaisussa: Nevalainen, Terttu & Traugott, Elizabeth Closs (toim.) *The Oxford Handbook of the History of English*. Oxford Handbooks in Linguistics. Oxford: Oxford University Press, 134–144.

Heikkinen, Vesa, Voutilainen, Eero, Lauerma, Petri, Tiililä, Ulla & Lounela, Mikko (toim.). 2012. *Genreanalyysi – tekstilajitutkimuksen käytäntöä*. Kotimaisten kielten keskuksen verkkojulkaisuja 29. [verkkoaineisto]. [viitattu 7.2.2018]. Helsinki: Kotimaisten kielten keskus. Saatavissa: http://kaino.kotus.fi/www/verkkojulkaisut/julk29/.

Hilpert, Martin. 2011. Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics* 16:4, 435–461.

Kampstra, Peter. 2008. Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of Statistical Software* 28, Code Snippet 1. Saatavissa: http://www.jstatsoft.org/v28/c01/.

Kroch, Anthony, Santorini, Beatrice & Diertani, Ariel. 2016. *The Penn Parsed Corpus of Modern British English* (PPCMBE2). Department of Linguistics, University of Pennsylvania. CD-ROM, 2. p., versio 1. Saatavissa: http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCMBE2-RELEASE-1.

Lijffijt, Jefrey, Säily, Tanja & Nevalainen, Terttu. 2012. CEECing the baseline: Lexical stability and significant change in a historical corpus. Julkaisussa: Tyrkkö, Jukka, Kilpiö, Matti, Nevalainen, Terttu & Rissanen, Matti (toim.) *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources*. Studies in Variation, Contacts and Change in English 10. Helsinki: VARIENG. [verkkoaineisto]. [viitattu 7.2.2018]. Saatavissa: http://www.helsinki.fi/varieng/series/volumes/10/lijffijt_saily_nevalainen/.

Marttila, Ville. 2014. *Creating Digital Editions for Corpus Linguistics: The Case of Potage Dyvers, a Family of Six Middle English Recipe Collections*. Väitöskirja. Helsinki: Helsingin yliopisto. Saatavissa: http://urn.fi/URN:ISBN:978-951-51-0060-3.

McEnery, Tony & Hardie, Andrew. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press.

Metsämuuronen, Jari. 2011. *Tutkimuksen tekemisen perusteet ihmistieteissä: tutkijalaitos*. (E-kirjan 1. p.). Helsinki: International Methelp.

Nevalainen, Terttu & Raumolin-Brunberg, Helena. 2003. *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. Longman Linguistics Library. London: Pearson Education.

Romaine, Suzanne. 1982. *Socio-historical Linguistics: Its Status and Methodology*. Cambridge Studies in Linguistics. Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt, Rosenbach, Anette, Bresnan, Joan & Wolk, Christoph. 2014. Culturally conditioned language change? A multi-variate analysis of genitive constructions in ARCHER. Julkaisussa: Hundt, Marianne (toim.) *Late Modern English Syntax*. Cambridge: Cambridge University Press, 133–152.

Säily, Tanja. 2014. *Sociolinguistic Variation in English Derivational Productivity: Studies and Methods in Diachronic Corpus Linguistics*. Mémoires de la Société Néophilologique de Helsinki XCIV. Helsinki: Société Néophilologique. Saatavissa: http://urn.fi/URN:ISBN:978-951-9040-50-9.

Vartiainen, Turo, Säily, Tanja & Hakala, Mikko. 2013. Variation in pronoun frequencies in early English letters: Gender-based or relationship-based? Julkaisussa: Tyrkkö, Jukka, Timofeeva, Olga & Salenius, Maria (toim.) *Ex Philologia Lux: Essays in Honour of Leena Kahlas-Tarkka*. Mémoires de la Société Néophilologique de Helsinki XC. Helsinki: Société Néophilologique, 233–255.
Ware, Colin. 2004. *Information Visualization: Perception for Design*. (2. p.). San Francisco: Morgan Kaufmann.